Supplemental Data

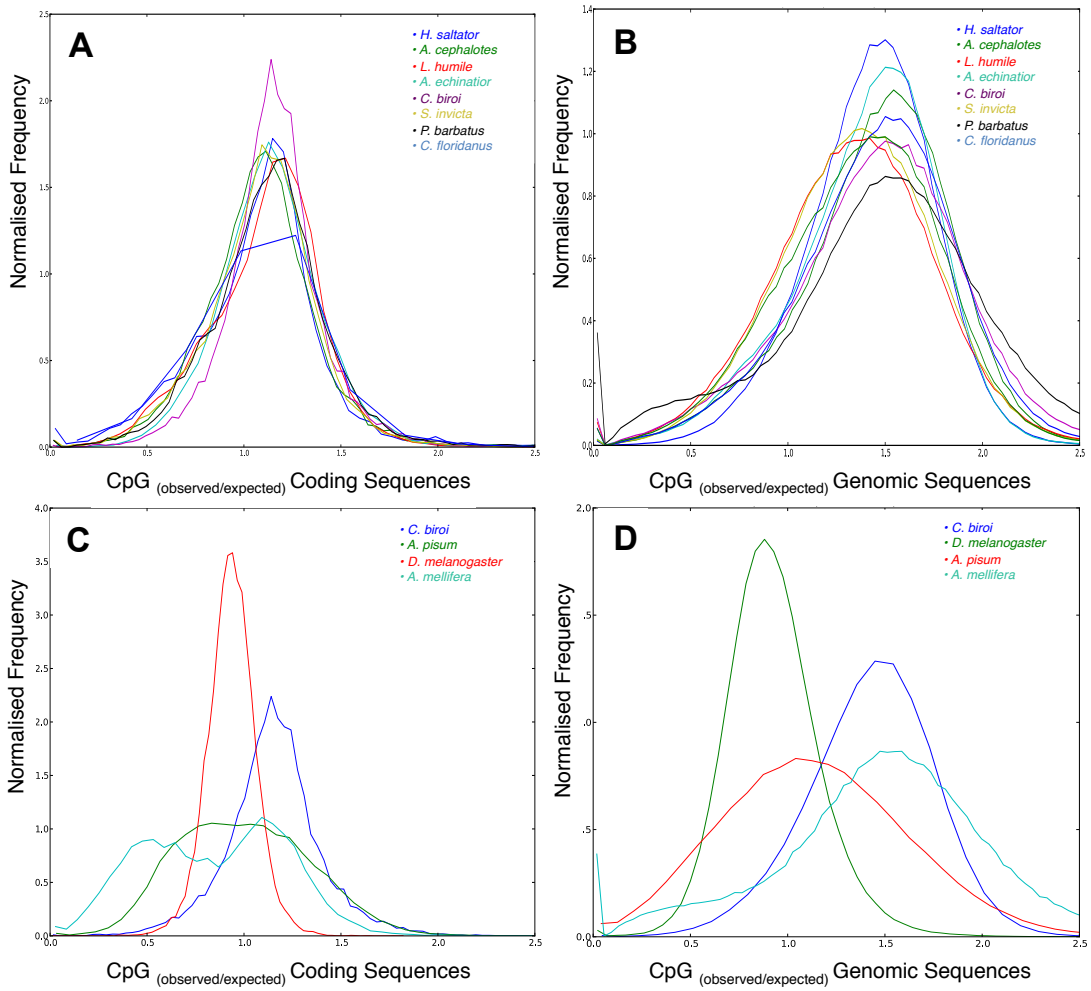


Figure S1 related to Table 1. Frequency plots showing $CpG_{(observed/expected)}$ ratios of annotated genes for sequenced ants (**A** and **B**) and other model species (**C** and **D**). **A** and **C)** $CpG_{(obs/exp)}$ ratios were calculated for the coding sequences of all annotated genes. **B** and **D)** The $CpG_{(obs/exp)}$ ratio for the entire genomic sequence of each species. The average $CpG_{(obs/exp)}$ ratio for the *C. biroi* genome assembly is 1.49. *Cerapachys biroi* Official Gene Set (OGS) 1.8 contains orthologs of all the genes required for DNA methylation (See also Supplemental Methods: DNA Methylation and Histone Modification).

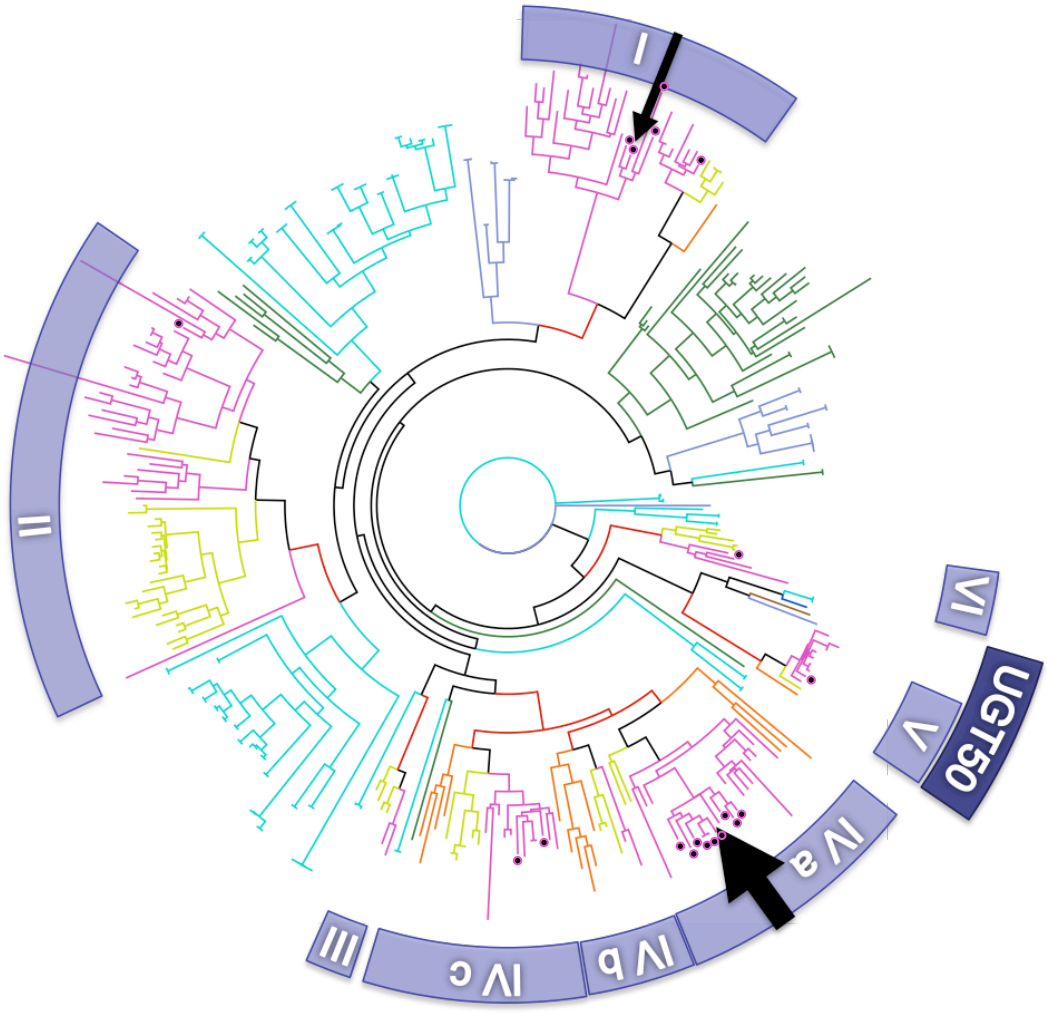


Figure S2 related to Table 1: Maximum likelihood phylogram of insect UGTs. Circles indicate *C. biroi* UGTs, while arrows indicate *C. biroi*-specific expansions (width of arrow is proportional to size of expansion). Roman numerals indicate hymenopteran-specific clades referred to in Supplemental Experimental Procedures: UDP Glycosyltransferases. Colors indicate taxa: pink – ants (*Harpegnathos saltator, Cerapachys biroi, Linepithema humile, Camponotus floridanus, Pogonomyrmex barbatus, Solenopsis invicta, Acromyrmex echinatior,* and *Atta cephalotes*), yellow – bees (*Apis mellifera* and *Bombus terrestris*), orange – *Nasonia vitripennis*, brown – *Tribolium castaneum*, light blue – *Bombyx mori*, green - *Acyrthosiphon pisum*, purple – *Drosophila melanogaster*. Hymenopteran clades are highlighted by red internal branches. Sequences and tree files have been deposited in the Dryad database (DOI pending).

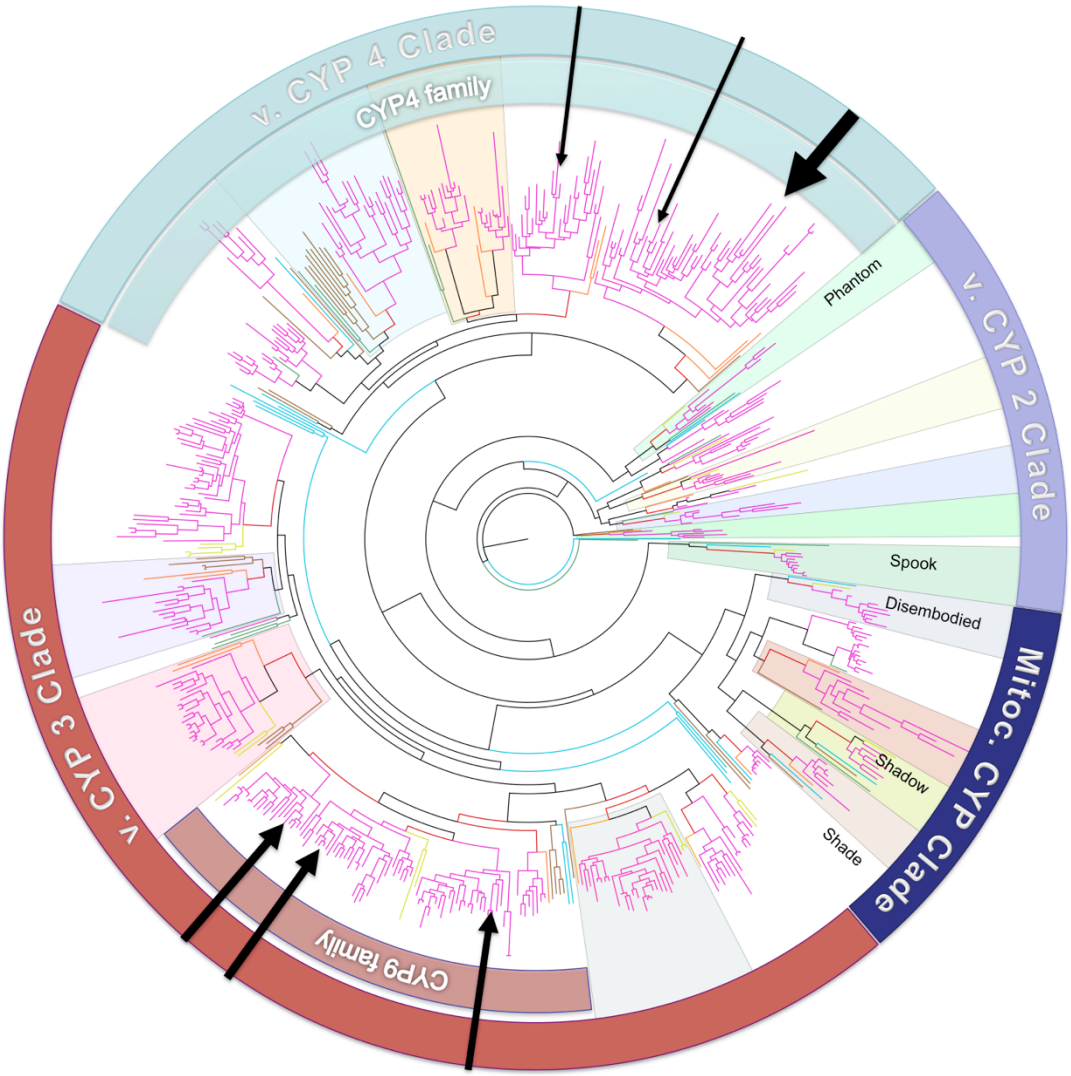


Figure S3 related to Table 1: Phylogeny of 573 insect cytochrome P450 genes. Orthologous vertebrate CYP clades are indicated on the outside ring. Arrows indicate *C. biroi*-specific expansions (width of arrow is proportional to size of expansion). Genes in the Halloween series are indicated by name. Colors indicate taxa as in Figure S2. Sequences and tree files have been deposited in the Dryad database (DOI pending). For methods see Supplemental Methods: Cytochrome P450 Genes.

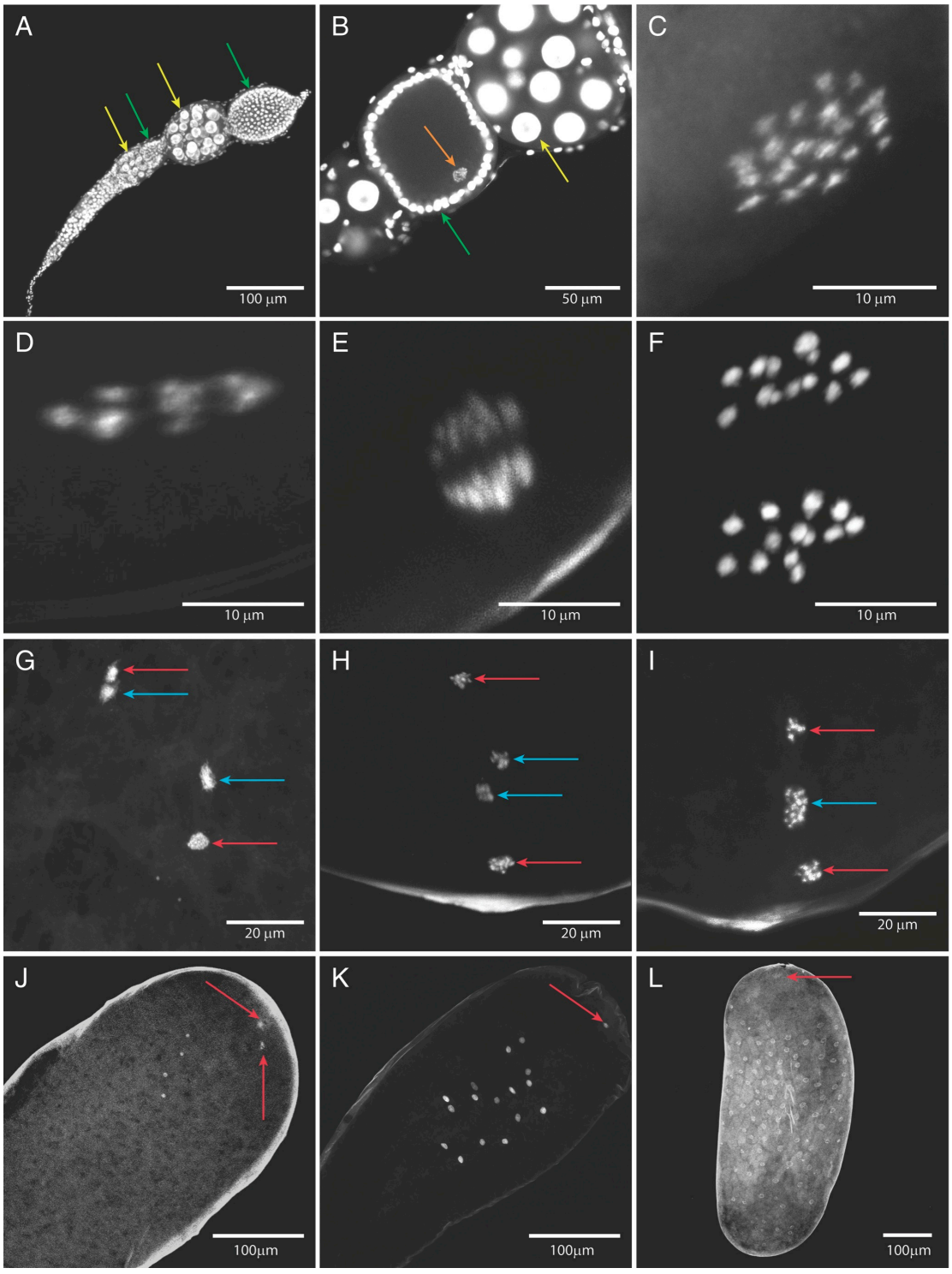

A
100 µm
B
50 µm
C
10 µm
D
10 µm
E
10 µm
F
10 µm
G
20 µm
H
20 µm
I
20 µm
J
100µm
K
100µm
L
100µm

Figure S4 related to Figure 2: Three-dimensional projections of DAPI-stained chromosomes in ovaries and eggs, showing automixis with central fusion. **(A-B)** Ovarioles. **(C-I)** Meiosis in < 2 hr old eggs (posterior egg boundary located toward bottom of each panel). **(J-L)** Mitosis and polar bodies in early embryos (egg posterior pole at top of panels). **A)** 3D projection of an ovariole with clusters of nurse cells (yellow arrows), and developing oocytes surrounded by follicular cells (green arrows). **B)** Optical section of an ovariole showing developing oocyte with a diploid nucleus (2n = 28) (orange arrow). Green arrow indicates one of the follicular cell nuclei, yellow arrow indicates one of the nurse cell nuclei. **C)** Diploid egg nucleus immediately post partum. **D)** Metaphase I, a few minutes post partum, with chromosomes aligned along the equatorial plane. **E)** Anaphase I, within 30 minutes post partum, showing separation of the homologous chromosomes. **F)** The two haploid nuclei (n = 14) after reductional division. **G)** Equational division of haploid nuclei, within one hour post partum. Nuclei indicated by red arrows are destined to become polar bodies. Non-homologous nuclei indicated by blue arrows will fuse. **H)** Migration of non-homologous nuclei (blue arrows) towards one another. **I)** Fusion of central products of meiosis to form a diploid nucleus (blue arrow), with two haploid polar bodies remaining (red arrows). Fusion occurs within one hour post partum. **J)** Embryo after two mitotic divisions of fused diploid nucleus (showing four diploid nuclei), within two hours post partum. Polar bodies (red arrows) have begun migrating towards the posterior pole of the egg. **K)** Embryo following four mitotic divisions, showing 16 diploid nuclei, and the fused polar body (red arrow). **L)** Embryo following seven mitotic divisions, showing 128 diploid nuclei and a degenerated fused polar body (red arrow).

Table S1 related to Table 1: Manually annotated chemosensory proteins in Hymenoptera, showing numbers of odorant receptor (OR), gustatory receptor (GR) and ionotropic receptor (IR) genes in each species with manually annotated olfactory genes. The first number in each column indicates putatively functional gene numbers, the second number indicates the total number of genes identified. Sources: [S22-S27, S29, S30]. NA = not manually annotated. Details of the exon structure, length, nucleotide sequence, position on genomic scaffold, and subfamily *sensu* Zhou *et al*. [S24] are deposited in the Dryad database (DOI pending).

	ORs	**GRs**	**IRs**	**OBPs**	**CSPs**
N. vitripennis	225/301	47/58	10	90	10
A. mellifera	163/174	10/13	10	21	6
H. saltator	347/377	17/21	23	NA	11
C. biroi	370/515	20/25	26	15	15
L. humile	337/367	97/117	32	12	14
C. floridanus	352/407	46/63	31	NA	12
P. barbatus	344/399	61/73	24	16	11

Supplemental Experimental Procedures

Cytology

Foraging colonies were subdivided into groups of 50 individuals, and the larvae removed to stimulate ovary activation. The first eggs to be observed were removed, then colonies observed at 10 min intervals, and all eggs produced during each interval were collected for preparation and staining.

Embryos were prepared according to [S1]: eggs were collected in phosphate buffered saline (PBS) pH 7.4 containing 1% Triton X-100 and boiled for 45 s then quenched on ice. Fixation was performed using 200 μl solution of 4% paraformaldehyde supplemented with 20 μl dimethyl sulfoxide (DMSO) and 600 μl n-heptane for 20 min at room temperature. Fixed embryos were washed three times with freezer-cold methanol and remaining chorion and vitelline membranes were removed with a fine brush. Embryos were sequentially hydrated in PBS and stained with 4',6-diamidino-2-phenylindole (DAPI). Images were acquired using a Zeiss LSM confocal microscope. 3D projections were constructed using Image J.

Sequencing

Genome assembly was achieved by combining data from Illumina HiSeq 2000 and Roche 454 platforms. For Illumina sequencing, five paired-end libraries (insert sizes of 200bp, 500bp, 800bp, 2kb, 5kb) were constructed, each from DNA from 75-150 pooled workers from multi-locus lineage 4 (MLL4; [S2]), resulting in 33 Gb of raw reads:

Table of raw sequencing read statistics. "Coverage depth" is calculated based on the assembled genome size (214 Mb).

Library ID	Average insert size (bp)	Average read length	Reads (M)	Bases (Mb)	Coverage depth
CERxbyDACDCAAPEI-2	164	90	75.98	6,838.4122	31.96
CERxbyDACDIAAPEI-3	444	90	110.53	9,947.5058	46.48
CERxbyDABDMAAPEI-4	732	90	77.91	7,012.2335	32.77
CERxbyDAADWAAPEI-2	2,275	49	121.73	5,964.9019	27.87
CERxbyDACDLAAPEI-1	4,981	49	61.76	3,026.2344	14.14
Total			447.91	32,789.2878	153.22

Reads were filtered for poly-A bases, ambiguous or poor quality sequences, adapter contamination and PCR duplicates:

Table of sequencing read statistics after filtering. "Coverage depth" is calculated based on the assembled genome size (214 Mb).

Library ID	Average insert size (bp)	Average read length	Usable reads (M)	Usable bases (Mb)	Coverage depth
CERxbyDACDCAAPEI-2	164	80	72.58	5,806.15184	27.08
CERxbyDACDIAAPEI-3	444	80	103.29	8,262.56624	38.54
CERxbyDABDMAAPEI-4	732	80	67.09	5,367.54432	25.04
CERxbyDAADWAAPEI-2	2,275	44	88.02	3,872.93474	18.07
CERxbyDACDLAAPEI-1	4,981	44	50.69	2,230.40048	10.40
total			381.67	25,539.59762	119.14

For Roche 454 sequencing, DNA was extracted from MLL1 and MLL6 (75 pooled workers each). Libraries constructed from each lineage were loaded onto half a PicoTiterPlate, and a titration run was performed to assess library quality. Titration plus full sequencing provided 526 Mb of raw reads (2.5x coverage).

To assist genome annotation, a 200bp transcriptome library was constructed from pooled RNA from all life stages: Adult callow and older workers in both reproductive and brood care phases, early and late stage pupae, early, middle and late stage larvae, and eggs. Approximately 75 individuals from each stage contributed to the RNA pool.

DNA was isolated using the Qiagen Genomic-tip 20/G extraction kit. RNA was isolated using a Trizol protocol followed by Qiagen RNeasy column purification. Libraries were prepared using manufacturer recommended protocols.

Assembly

SOAPdenovo [S3] was used to assemble the genome in three main steps. First, short insert library reads were split into k-mers and used to construct the de Bruijn graph. The graph was simplified according to default parameters and the k-mer path used to construct the contigs. Second, all usable reads were realigned onto the contigs, and the number of paired-end relationships between each pair of contigs was weighted and used to construct the scaffolds. Third, sequencing gaps in the scaffolds were closed by local assembly of 454 reads and Illumina paired-ends which mapped to scaffold gaps. These steps produced an assembly with an N50 scaffold length of 1,291,492bp and a total length of 214,372,378bp:

Table of genome assembly statistics.

	Contigs		Scaffolds	
	Size (bp)	**Number**	**Size (bp)**	**Number**
N90	6,347	7334	97,864	282
N80	11,611	4954	305,613	147
N70	17,430	3501	607,360	99
N60	24,125	2492	868,236	69
N50	31,934	1742	1,291,492	49
Longest	286,413		4,675,909	
Total Size	207,179,051		214,372,378	

Annotation

Gene predictions were generated using homology, *de novo* prediction and transcriptome analysis, and integrated using GLEAN [S4]. First, core genes were predicted using CEGMA [S5]. These were used to train SNAP [S6] *ab initio* gene prediction in the MAKER pipeline [S7], which further incorporated evidence from BLASTX [S8] and Exonerate [S9] (using *A. echinatior* Official Gene Set (OGS) 3.8, *C. floridanus* OGS 3.3 and *H. saltator* OGS 3.3), and EST evidence from *L. niger* and *S. invicta*. This produced a gene set consisting of 14,190 genes.

A separate homology search was performed against the NCBI non-redundant protein database using TBLASTN (E-value < $1e^{-5}$). Results were filtered for the most similar matches containing homologous regions ≥ 50% of the query protein. Genewise [S10] protein-nuclear alignment of *A. echinatior*, *A. mellifera*, *D. melanogaster*, and *Caenorhabditis elegans* was used to generate additional homology-based gene models. The union of all homology gene sets was used to create the final homology gene set, with the longest gene model from each overlapping set of genes identified selected as the model for the final set.

Augustus [S11] and SNAP [S6] were used for *de novo* gene prediction, after training on 1,556 intact genes from a *L. humile* homology search (carried out as described above). GLEAN used the intersection of both Augustus and SNAP gene sets to create a single *de novo* gene set. When gene models overlapped, the longer of the two models was chosen for the final set.

Using the *C. biroi* transcriptome data, Cufflinks [S12] was used to assemble 56,902 transcripts. Combining alternatively spliced transcripts and removing incomplete transcripts led to 8,022 non-redundant genes with complete ORFs.

The homology-based and *de novo* gene sets were integrated in GLEAN to generate a consensus gene set of 17,679 genes. We then compared the GLEAN gene set to the Cufflinks transcripts and MAKER gene set to identify overlapping models. The best gene model for each overlapping gene was incorporated into the GLEAN set, and all non-overlapping MAKER gene predictions with transcriptome coverage ≥ 1 FPKM were added to the gene set. Genes annotated as transposons, or containing no EST support nor

Swiss-Prot annotation, were removed from the gene set, resulting in a gene set containing 16,852 genes, 14,364 (85.24%) of which contain complete open reading frames. After manual annotation of the chemosensory proteins (detailed below), the *C. biroi* OGS 1.8 contained 17,263 genes, and was used for all work reported in this paper.

Table of gene annotation statistics categorized by method.

Gene set		**Number**	**Average transcript length (bp)**	**Average CDS length (bp)**	**Average exons per gene**	**Average exon length (bp)**	**Average intron length (bp)**
De novo	Augustus	11,961	5,132	1,522	5.55	274	794
	SNAP	33,543	4,677	836	3.85	217	1,349
	Merged	11,958	5,132	1,522	5.55	274	794
Homolog	*A. echinatior*	18,272	2,810	1,052	4.09	257	568
	L. humile	19,333	2,305	1,006	3.76	267	470
	A. mellifera	10,651	5,162	1,411	5.43	260	846
	D. melanogaster	7,963	4,127	1,218	4.94	246	738
	C. elegans	5,921	2,521	966	3.79	255	558
Transcripts with complete ORF		14,364	3,289	1,434	5.43	264	419
Homolog / *de novo* Consensus		17,730	3,535	1,071	4.01	267	820
MAKER		14,168	5,568	1,436	6.32	227	776
Final gene set (OGS 1.8)		17,263	4,740	1,224	4.75	257	937

To assess the annotation quality and completeness, we searched OGS 1.8 for 248 eukaryotic genes (the CEGMA gene set) conserved across *Arabidopsis thaliana, C. elegans, D. melanogaster, Homo sapiens, Saccharomyces cerevisiae and Schizosaccharomyces pombe* [S13]. Our gene set contained 247 core eukaryotic genes (99.6%), with 220 showing at least 80% sequence length overlap with their orthologous reference.

Functional Annotation

Gene functions were assigned by BLASTP to the Swiss-Prot database [S14], with each query/target match requiring at least 50% identity. Gene motifs and domains were determined by InterProScan (version 4.3 [S15]) against 10 models (blastProDom, FPrintScan, HMMPfam, HMMPIR, HMMPanther, HMMTigr, HMMSmart, SuperFamily, Gene3D, and ProfileScan (InterPro release 36.0)). Gene ontology (GO) assignment was obtained from both InterProScan results and from protein domain – GO associations [S16]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation was done through the KEGG Automatic Annotation Server [S17], using the 'single direction best hit' method (www.genome.jp/kaas-bin/kaas_main).

Table of functional annotation statistics.

Functional database	**Number of Genes Annotated**
Swiss-Prot	9,295 (55.2%)
InterPro	9,628 (57.1%)
GO	7,835 (46.5%)
KEGG	5,212 (30.9%)

Gene functions were assigned for all eight ant species with currently sequenced genomes, and grouped according to their InterPro classification calculated above. Using Dixon's Q-test, InterPro classifications in which *C. biroi* gene number was significantly under- or over-represented were identified:

Table showing gene-family sizes based on IPR annotation for the eight sequenced ant genomes. Families shown below are those in which *C. biroi* has a significant expansion or contraction. Hs – *H. saltator,* Cb – *C. biroi*, Lh – *L. humile,* Cf – *C. floridanus*, Pb – *P. barbatus*, Si – *S. invicta,* Ac – *A. cephalotes*, Ae – *A. echinatior*.

Hs	**Cb**	**Lh**	**Cf**	**Pb**	**Si**	**Ac**	**Ae**	**Description**
1	7	1	1	1	1	1	1	IPR009602, FAM92 protein
2	7	1	2	2	1	2	2	IPR001087, Lipase, GDSL
3	12	3	3	3	3	3	3	IPR002125, CMP/dCMP deaminase, zinc-binding
8	18	8	7	7	7	9	8	IPR001763, Rhodanese-like
25	34	26	28	26	27	26	25	IPR001394, Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2
28	31	15	23	15	22	14	19	IPR002213, UDP-glucuronosyl/UDP-glucosyltransferase
1	7	1	1	1	1	1	1	IPR008710, Nicastrin
11	17	6	8	2	2	1	1	IPR022083, KIF-1 binding protein
1	22	20	0	25	3	1	19	IPR003595, Protein-tyrosine phosphatase, catalytic
2	11	2	8	2	4	5	2	IPR021190, Peptidase M10A, matrix metallopeptidase
20	26	19	18	19	20	21	14	IPR017981, GPCR, family 2-like
13	20	15	14	17	18	16	12	IPR000832, GPCR, family 2, secretin-like
8	31	23	9	18	9	11	22	IPR009053, Prefoldin
2	12	2	10	2	4	9	2	IPR000585, Hemopexin/matrixin
42	81	81	42	22	44	27	27	IPR006612, Zinc finger, C2CH-type

Table showing gene-family sizes based on IPR annotation (continued).

Hs	Cb	Lh	Cf	Pb	Si	Ac	Ae	Description
11	3	6	13	4	16	15	5	IPR017956, AT hook, DNA-binding motif
4	28	32	7	31	5	7	20	IPR002041, Ran GTPase
5	15	7	12	9	10	10	7	IPR006026, Peptidase, metallopeptidase
6	12	10	8	10	9	10	8	IPR000337, GPCR, family 3
3	13	13	3	9	3	5	8	IPR006687, Small GTPase superfamily, SAR1-type
1	18	7	4	4	10	3	3	IPR007527, Zinc finger, SWIM-type
115	67	87	76	85	75	74	86	IPR009003, Peptidase cysteine/serine, trypsin-like
39	34	46	41	39	40	43	50	IPR011042, Six-bladed beta-propeller, TolB-like
2	12	2	11	2	5	6	2	IPR001818, Peptidase M10, metallopeptidase
15	20	21	15	20	10	16	16	IPR024156, Small GTPase superfamily, ARF type

Table showing Swiss-Prot annotations in which *C. biroi* shows significant expansion or contraction. For species abbreviations refer to IPR table above.

Hs	Cb	Lh	Cf	Pb	Si	Ac	Ae	Description
1	0	1	1	1	1	1	1	Protein shifted
1	8	1	2	1	2	1	1	Valacyclovir hydrolase
1	7	1	1	1	0	1	1	Deoxycytidylate deaminase

Phylogeny Reconstruction and Gene Expansions

To gain insight into the evolution of ant gene families, we clustered genes of ten Hymenoptera species (*A. echinatior, A. mellifera, A. cephalotes, C. floridanus, C. biroi, H. saltator, L. humile, N. vitripennis, P. barbatus, S. invicta)* and the outgroup *D. melanogaster* into gene families using OrthoMCL [S18]. The longest isoform of each gene from each species was used in all against all BLASTP with an E-value cutoff of 1×10^{-5}. An inflation parameter of 1.5 was used for gene clustering.

After constructing gene families, the program CAFE [S19] was used to identify gene families that have undergone expansion or contraction in the eight ant species. After manual checking of gene annotations, only Twist (IPR015789) was found to be significantly over-represented in *C. biroi*. Twist is involved in gastrulation and differentiation of adult musculature, and may therefore play a role in the development of morphological characteristics of *C. biroi*.

Table showing gene families expanded in *C. biroi* identified with CAFE and verified with manual annotation. For species abbreviations refer to IPR table above.

Hs	Cb	Lh	Cf	Pb	Si	Ac	Ae	IPR
2	3	1	1	1	1	3	1	IPR015789; Twist.

To construct the phylogeny of these 11 species, 3,164 genes belonging to single-gene gene families with orthologs in all 11 species were aligned using MUSCLE [S20] and concatenated into a supergene for each species. A maximum likelihood phylogeny was constructed using PhyML [S21], with the JTT+G+I model for amino acid substitution. The phylogeny was rooted with *D. melanogaster*. Boostrap support for all nodes was 100% (100 replicates). The phylogeny is given in Figure 5.

Cerapachys biroi-Specific Genes

Genes present in *C. biroi* but containing no orthologs in the seven other ant genomes were considered *C. biroi* specific. All *C. biroi*-specific genes were tested for GO term, KEGG and IPR enrichment (see following tables) (using a FDR q-value of 0.05 for each). KEGG Ontology K00699 (UGT; glucuronosyltransferase [EC:2.4.1.17]) was found in ten enriched KEGG pathways, while K07424 (CYP3A; cytochrome P450, family 3, subfamily A [EC:1.14.14.1]) was found in seven enriched pathways. These pathways are primarily involved in cytochrome-based metabolism, and lipid and carbohydrate metabolism. By clustering enriched GO terms into parent-child relationships, *C. biroi* was shown to have only eight enriched groups, which are predominantly involved in DNA metabolism, cytochrome-based metabolism, and odor perception. We therefore proceeded to manually annotate the chemosensory protein families, the UDP glycosyltransferase superfamily and the cytochrome P450 superfamily – all of which showed significant expansion in *C. biroi* (see tables above and below) and are involved in a wide variety of behavioral and metabolic processes.

Table showing GO enrichment in *C. biroi*-specific genes. GO terms have been organized according to parent-child relationships. All GO terms that are within the same parent-child tree are highlighted in identical colors, and all related GO hierarchies are within the same border. Class represents Molecular Function (MF) or Biological Process (BP).

GO ID	GO Term	Class	Level	P-value	Gene count
GO:0046983	protein dimerization activity	MF	4	2.54E-15	23
GO:0016485	protein processing	BP	6	4.76E-07	9
GO:0006950	response to stress	BP	3	3.64E-59	23
GO:0016779	nucleotidyltransferase activity	MF	5	0.004683918	15
GO:0016788	hydrolase activity, acting on ester bonds	MF	4	2.37E-35	45
GO:0006259	DNA metabolic process	BP	5	1.19E-82	56
GO:0034061	DNA polymerase activity	MF	6	1.45E-06	14
GO:0004518	nuclease activity	MF	5	6.66E-87	33
GO:0006310	DNA recombination	BP	6	9.24E-05	9
GO:0015074	DNA integration	BP	6	4.88E-15	25
GO:0006281	DNA repair	BP	6	1.04E-75	19
GO:0003887	DNA-directed DNA polymerase activity	MF	7	2.55E-06	13
GO:0004523	ribonuclease H activity	MF	9	0.00089136	4
GO:0008408	3'-5' exonuclease activity	MF	7	6.24E-06	12
GO:0003676	nucleic acid binding	MF	3	0.000285676	137
GO:0032501	multicellular organismal process	BP	2	8.53E-19	54
GO:0004872	receptor activity	MF	4	8.31E-12	60
GO:0004871	signal transducer activity	MF	3	1.76E-09	61
GO:0007608	sensory perception of smell	BP	7	1.57E-30	53
GO:0004888	transmembrane receptor activity	MF	5	2.16E-14	59
GO:0004930	G-protein coupled receptor activity	MF	6	6.90E-23	58
GO:0004984	olfactory receptor activity	MF	7	1.57E-30	53
GO:0005549	odorant binding	MF	3	5.54E-28	53
GO:0009055	electron carrier activity	MF	2	1.27E-18	36
GO:0016491	oxidoreductase activity	MF	3	0.000188573	56
GO:0004497	monooxygenase activity	MF	4	9.03E-26	37
GO:0016758	transferase activity, transferring hexosyl groups	MF	5	2.08E-05	16
GO:0005506	iron ion binding	MF	7	3.62E-18	40
GO:0020037	heme binding	MF	4	1.31E-21	37

Table showing non-redundant GO term enrichment of *C. biroi*-specific genes. All genes represented more than once in the previous table were assigned to their highest level GO term only. GO terms have been organized according to parent-child relationships. All GO terms that are within the same parent-child tree are highlighted in identical colors, and all related GO hierarchies are within the same border. Class represents Molecular Function (MF) or Biological Process (BP).

GO ID	GO Term	Class	Level	P-value	Gene count
GO:0046983	protein dimerization activity	MF	4	2.54E-15	23
GO:0016485	protein processing	BP	6	4.76E-07	9
GO:0006950	response to stress	BP	3	3.64E-59	23
GO:0016779	nucleotidyltransferase activity	MF	5	0.004683918	15
GO:0016788	hydrolase activity, acting on ester bonds	MF	4	2.37E-35	45
GO:0034061	DNA polymerase activity	MF	6	1.45E-06	14
GO:0015074	DNA integration	BP	6	4.88E-15	25
GO:0006281	DNA repair	BP	6	1.04E-75	19
GO:0003887	DNA-directed DNA polymerase activity	MF	7	2.55E-06	13
GO:0004523	ribonuclease H activity	MF	9	0.00089136	4
GO:0008408	3'-5' exonuclease activity	MF	7	6.24E-06	12
GO:0003676	nucleic acid binding	MF	3	0.000285676	137
GO:0032501	multicellular organismal process	BP	2	8.53E-19	54
GO:0004872	receptor activity	MF	4	8.31E-12	60
GO:0004871	signal transducer activity	MF	3	1.76E-09	61
GO:0007608	sensory perception of smell	BP	7	1.57E-30	53
GO:0004888	transmembrane receptor activity	MF	5	2.16E-14	59
GO:0004930	G-protein coupled receptor activity	MF	6	6.90E-23	58
GO:0016491	oxidoreductase activity	MF	3	0.000188573	56
GO:0016758	transferase activity, transferring hexosyl groups	MF	5	2.08E-05	16
GO:0005506	iron ion binding	MF	7	3.62E-18	40

Table of IPR enrichment statistics for *C. biroi*-specific genes.

IPR ID	IPR Title	P-value	Number of genes
IPR000305	Excinuclease ABC, C subunit, N-terminal	2.54E-133	67
IPR012337	Ribonuclease H-like	1.24E-41	65
IPR001128	Cytochrome P450	1.70E-30	36
IPR004117	Olfactory receptor, *Drosophila*	1.26E-28	53
IPR006612	Zinc finger, C2CH-type	5.51E-24	31
IPR008906	HAT dimerisation	1.55E-14	24
IPR008710	Nicastrin	9.07E-10	9
IPR015517	Cytidine deaminase	9.07E-10	9
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	1.80E-09	15
IPR000794	Beta-ketoacyl synthase	1.84E-09	13
IPR002125	CMP/dCMP deaminase, zinc-binding	1.73E-08	9
IPR010285	DNA helicase PIF1, ATP-dependent	8.78E-08	10
IPR016193	Cytidine deaminase-like	1.36E-07	9
IPR004868	DNA-directed DNA polymerase, family B, mitochondria/virus	1.82E-07	11
IPR002403	Cytochrome P450, E-class, group IV	1.15E-06	8
IPR009602	Protein of unknown function DUF1208	1.40E-06	6
IPR001763	Rhodanese-like	2.49E-06	9
IPR005312	Protein of unknown function DUF1759	7.73E-06	14
IPR022083	KIF-1 binding protein C-terminal	1.63E-05	8
IPR023211	DNA polymerase, palm domain	2.94E-05	9
IPR000073	Alpha/beta hydrolase fold-1	3.83E-05	10
IPR004211	Recombination endonuclease VII	0.000103	8
IPR013196	Helix-turn-helix, type 11	0.000169	4
IPR016473	dCMP deaminase	0.000169	4
IPR017446	Polyprenyl synthetase-related	0.000268	5
IPR015569	Peptidase M1, aminopeptidase N	0.000373	6
IPR002156	Ribonuclease H domain	0.000475	4
IPR002401	Cytochrome P450, E-class, group I	0.000852	12
IPR008949	Terpenoid synthase	0.000863	5
IPR001930	Peptidase M1, alanine aminopeptidase/leukotriene A4 hydrolase	0.00112	10
IPR006047	Glycosyl hydrolase, family 13, catalytic domain	0.00139	5

Table of IPR enrichment statistics for *C. biroi*-specific genes (continued).

IPR ID	IPR Title	P-value	Number of genes
IPR015902	Alpha amylase	0.0014	5
IPR020064	ABC transporter, G1-like	0.0014	5
IPR010562	Haemolymph juvenile hormone binding	0.0018	7
IPR014782	Peptidase M1, membrane alanine aminopeptidase, N-terminal	0.0030	9
IPR000092	Polyprenyl synthetase	0.0033	4
IPR004875	DDE superfamily endonuclease, CENP-B-like	0.0033	4
IPR006096	Glutamate/phenylalanine/leucine/valine dehydrogenase, C-terminal	0.0041	3
IPR005055	Insect pheromone-binding protein A10/OS-D	0.0059	5

Table of KEGG pathway enrichment statistics for *C. biroi*-specific genes.

Map ID	Map Title	P-value	Number of genes
map00310	Lysine degradation	2.64E-30	45
map00140	Steroid hormone biosynthesis	8.47E-13	18
map00980	Metabolism of xenobiotics by cytochrome P450	7.95E-12	20
map00982	Drug metabolism - cytochrome P450	5.25E-11	19
map00830	Retinol metabolism	5.98E-11	18
map00983	Drug metabolism - other enzymes	1.07E-09	18
map00500	Starch and sucrose metabolism	4.36E-09	18
map00061	Fatty acid biosynthesis	9.13E-08	11
map00053	Ascorbate and aldarate metabolism	7.36E-07	12
map00514	Other types of O-glycan biosynthesis	1.03E-06	13
map00040	Pentose and glucuronate interconversions	5.26E-06	12
map00860	Porphyrin and chlorophyll metabolism	3.25E-05	12
map00591	Linoleic acid metabolism	8.29E-05	7
map00627	Aminobenzoate degradation	0.00026	7
map04910	Insulin signaling pathway	0.0023	14

Chemoreception

Ants exhibit perhaps the most complex social coordination of all invertebrates, and the large expansions of chemoreceptor genes observed in ant genomes are hypothesized to contribute to this coordination by facilitating chemical communication [S22-S24]. However, gene families involved in olfaction are notoriously difficult for automatic annotation pipelines due to their rapid evolutionary rates and low level of expression reflected in whole-body EST libraries. In order to facilitate future genomic studies of communication in *C. biroi*, we undertook an exhaustive manual annotation of the three receptor families and two small soluble protein families implicated in olfaction in insects, namely the odorant receptors (ORs), gustatory receptors (GRs), ionotropic receptors (IRs), odorant binding proteins (OBPs), and chemosensory proteins (CSPs).

For each gene family we identified loci via TBLASTN homology searches of the whole genome with query sequences from *P. barbatus* [S23], *L. humile* [S22], and *A. mellifera* [S25-S27]. We then used TBLASTN to locate exons and the Apollo genome annotator to manually define exon boundaries [S8, S28]. We refined our annotations by using the MEGA alignment viewer with both manual and Muscle [S20] alignments to compare our annotations with homologous sequences. Roughly half of all olfactory genes could not be fully reconstructed due to gaps in the genomic sequence. These are indicated in the gene names following the format of CD Smith *et al*. [S22] and CR Smith *et al*. [S23].

Table of gene nomenclature for indicating incompletely annotated genes.

NTE	Missing sequence at N terminus
INT	Missing sequence in the middle of gene
CTE	Missing sequence at C terminus
NI	Missing N terminus and section in the middle of gene
NC	Missing N and C terminus
IC	Missing section in the middle of gene and C terminus

Especially for the OR subfamily, we classified many genes as probable nonfunctional pseudogenes based on premature stop codons, missing exons, frameshifts, and incorrect splice sites. Again, we followed the format of CD Smith *et al*. [S22] and CR Smith *et al*. [S23] in the indication of pseudogenization in the gene names, with the modification that we indicated cases in which genes might have been incorrectly identified as pseudogenes either because they may have non-canonical splice sites or because of potential assembly-introduced spurious frameshifts, e.g. due to homopolymer errors arising from the 454 data.

Table showing additional nomenclature for OR genes.

PSE	Pseudogene
P+N/I/C	Pseudogene and missing sequence
(F)	Could be functional with assembly-introduced false frameshift
(S)	Could be functional with non-canonical splice sites

We have included all manually annotated genes in the official gene set, making the *C. biroi* genome the first ant genome with manually curated olfactory genes in the official gene set.

Odorant receptors: manual annotation revealed 506 odorant receptor genes and gene fragments with at least 200 amino acid residues (approximately half of a full gene) (Table S1). 255 of these sequences were putatively functional full-length genes, and an additional 109 were putatively functional but the full sequence could not be reconstructed because of gaps in the genome sequence. The remaining 141 had premature stop codons, large deletions and entire exons missing, frameshifts, and/or missing splice sites. An additional 128 fragments with fewer than 200 amino acid residues were identified, and many of these likely represent additional OR genes. ORs were named by assigning them to OR subfamilies based on exon structure, ordering these subfamilies to maximize correspondence to CD Smith *et al*. [S22] and CR Smith *et al*. [S23], and then numbering the ORs sequentially starting with the universally conserved odorant co-receptor (ORCO) as *CbirOR1*.

Cerapachys biroi has more putatively functional ORs and pseudogenized ORs than any other insect species annotated to date. Including pseudogenes, *C. biroi* has 24-38% more ORs than the other ant species with manually annotated ORs, which in turn each have more ORs than any non-ant insect. The exceptionally high percentage of OR pseudogenes in *C. biroi* relative to other ants may indicate recent rapid expansions and dynamic evolution in this gene family in the *C. biroi* lineage. Ants are hypothesized to possess expanded olfactory gene repertoires to facilitate social communication [S24, S30], and like other ants, *C. biroi* has advanced chemical communication and may use pheromones to facilitate behaviors such as worker policing, reproductive coordination, and foraging activity [S31, S32]. The expansion of OR genes in *C. biroi* relative to other ants could be related to the fact that the species is entirely subterranean and workers are blind, probably relying even more heavily on their olfactory sense. As obligate myrmecophages with a broad prey spectrum, *C. biroi* foragers must be able to recognize a variety of ant species and locate their nests. The expansion of ORs and their rapid evolution may thus also indicate a specialization in "eavesdropping" on the recognition and orientation pheromones of other ant species, and the necessity to distinguish between own brood and prey items.

Gustatory and ionotropic receptors: The number of GRs and IRs in *C. biroi* is far more modest, falling in the low to middle range for ants. The 20 functional GRs in *C. biroi*, 17 in *H. saltator*, and ten in *A. mellifera* indicate that low GR copy number might have been the ancestral state in ants, with the copy number having roughly doubled in the rest of the formicoids after the dorylomorphs split off. The number of IRs in *C. biroi* (26) is comparable to the number in other ant species, which have from 23 to 32 IRs with no clear phylogenetic signal in copy number evolution.

Odorant binding and chemosensory proteins: These two families of small soluble proteins are involved in a variety of non-chemosensory physiological processes [S33-S36]; nevertheless, some OBPs and CSPs appear to be essential for olfaction (reviewed in [S37, S38]). We found 15 full-length OBPs and 15 full-length CSPs, named for homology to *A. mellifera* proteins where single copy orthology exists, and numbered arbitrarily when in paralogous expansions. These copy numbers are comparable to other ant species (12-16 OBPs, 12-21 CSPs). However, ant OBP copy numbers are low relative to non-aculeate neopterans (41-90 OBPs) [S8, S33].

UDP Glycosyltransferases (UGTs)

UDP glycosyltransferases compose a superfamily of proteins found in animals, plants, bacteria and viruses. In insects, UGTs are used to solubilize endogenous and exogenous compounds, altering their bioactivity and/or allowing the compound to be excreted [S39]. UGTs play an important role in insecticide resistance [S40], and have also been implicated in odor perception [S41, S42].

NCBI was searched for UDP glycosyltransferases and UGT-like proteins (such as UDP glucuronyltransferases) from the genomes of the ants *H. saltator, L. humile, C. floridanus, P. barbatus, S. invicta, A. echinatior,* and *A. cephalotes*, the bees *A. mellifera* and *B. terrestris*, the wasp *N. vitripennis*, the beetle *T. castaneum*, the silkworm *B. mori*, the pea aphid *A. pisum*, and the vinegar fly *D. melanogaster*. A total of 244 protein sequences were downloaded.

To identify as many UGT sequences in the ants as possible, we arbitrarily subdivided all UGT sequences into subgroups, used the subgroups to construct sequence models, and then identified all proteins that matched the models. All 244 previously annotated UGT protein sequences were aligned using MUSCLE (default parameters), and a neighbor joining (NJ) tree was constructed using Mega 5.0. The NJ tree was then subdivided manually into 21 clades with approximately equal within-clade divergence (each clade contained between four and 51 taxa).

Each of the 21 clades was used to generate a Hidden Markov Model using HMMER software, and the models were then used to search the peptide sequences of all species listed above, plus those of *C. biroi*, to identify UGT candidate sequences. After a preliminary assessment of results, all sequences with an alignment bit score < 120 were discarded.

The UDP binding site signature motif was obtained from Mackenzie *et al*. [S43], and all candidate sequences were assessed to see if they contained the signature. The motif and 100 flanking amino acids (if present) were aligned from all candidate genes that matched the signature sequence. The trimmed protein sequences were then realigned with the remaining unassigned proteins using ClustalW [S44], and any proteins matching > 90% of the signature motif were used to update the signature. The updated motif was then used to re-search the whole protein list, with new sequences trimmed, aligned and used to search the remaining sequences. This process was repeated until no new proteins matching the signature motif were found.

The final signature motif was then used to re-search all *C. biroi* peptides, but no genes additional to those found in the initial UGT signature search were found. The final set of *C. biroi* genes that matched the signature motif were then checked against the transcriptome data to ensure they were transcribed.

The signature motif and flanking 100 bp for the 321 identified genes were then aligned using MUSCLE, and a maximum likelihood tree constructed using Garli 2.0 (Poisson+G+I evolutionary model; best tree of five runs chosen) (Figure S2).

A total of 107 UGTs were identified in the eight ant species searched (sequence alignment and tree file have been deposited in the Dryad database. DOI pending). *Cerapachys biroi* has 21 UGT proteins, the largest number of any sequenced ant species.

Table of the number of UGT proteins identified in the sequenced ant species. For species abbreviations refer to IPR table above.

Hs	**Cb**	**Lh**	**Cf**	**Pb**	**Si**	**Ac**	**Ae**	**TOTAL**
18	21	10	13	11	15	10	9	107

The hymenopteran genes segregated into six distinct groups within the insect UGTs, with group IV indicating a pre-hymenopteran expansion into three distinct subgroups (IVa-c). *Cerapachys biroi* is represented in every group, but is not represented in subgroup IVb, which among the ant taxa only contains two genes from *H. saltator*. Groups I, II, III, and subgroups IVa and IVc each contain at least one representative from each of the eight ant species.

Group III and subgroup IVa have lineage-specific expansions in *C. biroi*. These two (sub-) groups also contain the only other two large lineage-specific ant expansions, involving *S. invicta* (subgroup IVa) and *C. floridanus* (group III). In subgroup IVa, eight of the ten genes in the *C. biroi* expansion are found on the same scaffold. This region of tandem UGT duplication also contains partial UGT sequences and a few non-UGT proteins. Because these genes are found at the end of the scaffold, it cannot be ruled out that the additional two subgroup IVa genes are also located in tandem with the other eight genes.

Group II contains four separate ant clades, with each clade containing sequences from at least six ant species. All eight ant species are represented in group II, but *C. biroi* is only represented in one of the four clades.

Table of ant species representation among the six UGT families.

(Sub-) group	Number of UGT genes	Number of ant species	Number of *C. biroi* UGTs
I	9	8	1
II	39	8	1
III	25	8	4
IVa	22	8	10
IVb	2	1	0
IVc	12	8	2
V	2	2	1
VI	4	4	1

The signature motif for all ant UGTs identified is: [FIVY]-[FILMVW]-X-[HNQ]-[GS]-G-[AFILMGV]-X-[GS]-X-X-[EK]-[AGST]-[AFILV]-X-X-X-[AILV]-[PR]-[ILMTV]-[FILTV]-[AGILV]-[CFILMTV]-[PS]-[FILMNV]-[CFILMVY]-X-[DEN]-[HQ]

Sequences from family UGT50 from additional taxa [S45] revealed group V to be part of the UGT50 family. Just as all insects previously studied were represented within this family (with the exception of *A. pisum*) [S45], so are all ant species. This highly conserved family has been suggested to be homologous to mammalian UGT8 [S45], which conjugates galactose to a sphingosine, an important component of lipid bilayer formation and apoptotic signaling.

Possible functions: It is difficult to assign function to the ant UGTs, as the family is extremely diverse, binds a large range of substrates, and is poorly characterized in insects [S46]. The baculovirus ecdysteroid UGTs are most closely related to UGT33 in *B. mori*, which is sister to group II hymenopteran UGTs. If these UGTs play a role in ecdysteroid regulation, then most ants have acquired a set of four ecdysteroid UGT-like proteins, while *C. biroi* has maintained only a single copy. UGTs have been associated with olfaction in vertebrates [S47] and invertebrates [S42]. Antennal expression enriched bmUGT013829 in *B. mori* [S48] and UGT35a and UGT35b in *D. melanogaster* [S42] are not closely related to each other, and we did not find any hymenopteran UGTs that are closely related to either. The detoxification of substances by UGTs is well known [S43, S49, S50], and the greatest UGT expansions can be found in herbivorous insects that require extensive detoxification pathways. However, the myrmecophagous diet of *C. biroi* does not immediately suggest a need for a broad range of detoxification enzymes. A final possible role for UGTs in *C.*

biroi may be for sclerotisation of the cuticle [S51], yet it should be noted that the quercetin glycosylating UGT40 family (conferring UV protection to pupae in cocoons [S52]) has no closely related hymenopteran genes. In conclusion, the *C. biroi* genome contains many UGTs, possibly with important functions. However, these cannot be directly inferred from phylogeny, but await assignment through functional analysis.

Cytochrome P450 Genes

NCBI was searched for cytochrome P450 (CYP) protein sequences from the hymenopteran, lepidopteran and dipteran genomes used to construct the UGT phylogeny. A total of 892 non-ant CYP proteins were downloaded. To create a non-redundant protein list, proteins were clustered according to sequence similarity, and a single sequence from each cluster chosen as a representative for phylogenetic analysis. All 892 proteins were clustered using h-cd-hit [S53] (clustering parameters were 0.9 and 0.6 for the first and second cluster levels, respectively). H-cd-hit created 111 clusters. The representatives from each cluster were aligned using MUSCLE, and a NJ tree constructed using Mega 5.0. The NJ tree was then divided manually into 18 clades with approximately equal within-clade sequence divergence.

Each of the 18 clades was used to generate Hidden Markov Models using HMMER, and these models were used to search the peptide sequences of each of the sequenced ant species. All sequences with an alignment bit score < 120 were discarded. A total of 901 ant CYP proteins were identified.

The 901 ant CYPs and 111 representative insect CYPs were aligned using MUSCLE, with manual verification to ensure that all conserved motifs [S54] were present and correctly aligned.

26 ant CYP genes were found to be incorrectly annotated tandem duplications, which were subsequently re-annotated for the alignment. Sequences that were truncated, or missing either the active site consensus sequence or the K-helix conserved motif, were discarded. This led to approximately half of the sequences being discarded, leaving a final set of 573 sequences, including 462 ant CYPs (69 from *C. biroi*) and 111 non-redundant CYPs from other invertebrates.

The manually corrected alignment of the 573 sequences was further edited to remove the hypervariable non-conserved N-terminal up to (but not including) the N-terminal anchor sequence, as well as the region between the C-helix and I-helix. This alignment was then used to construct a maximum likelihood phylogeny using Garli 2.0 (Poisson+G+I evolutionary model; best tree of two runs chosen) (Figure S3; sequence alignment and tree file have been deposited in the Dryad database. DOI pending).

The insect CYPs correctly grouped into the four major clades identified by Feyereisen [S55]. *Cerapachys biroi*-specific gene expansions were found only in the CYP4 and CYP9 gene families. All CYPs from the Halloween series, which are essential for development, were found in *C. biroi* and most other ant species.

Like the UGT protein superfamily, cytochrome P450 genes are involved in a diverse number of functions. Many insects have expansions in clades 3 and 4, and some representatives from these clades are involved in xenobiotic metabolism [S55]. Some CYPs in clade 4 are also involved in odorant and pheromone metabolism [S55].

Table of cytochrome P450 genes identified in the eight sequenced ant species. The bottom row is the total number of CYPs identified for each ant species, and includes the genes listed separately for CYP4 and CYP9. For species abbreviations refer to IPR table above.

	Hs	**Cb**	**Lh**	**Cf**	**Pb**	**Si**	**Ac**	**Ae**
CYP4	12	21	23	29	17	21	9	14
CYP9	17	20	0	13	0	10	0	22
All CYPs	58	69	52	84	39	63	28	65

Vitellogenin Annotation and Phylogeny

Previously annotated *vitellogenin* (*Vg*) genes from each of the seven previously sequenced ant genomes and *Apis mellifera* were used to re-query all eight ant genomes using TBLASTN [65], PHMMER [66] and Exonerate [67]. Exon/intron boundaries were predicted using TBLASTN and Exonerate results, and manually refined in the Apollo genome annotator [68]. Annotations were refined using MUSCLE alignments and MEGA alignment viewer [69]. Gene nucleotide sequences were aligned by codons using the MUSCLE alignment algorithm, and Guidance [70] was used to identify ambiguously aligned codons. Third position nucleotides and low scoring alignment columns (Guidance score < 95) were removed. A maximum likelihood phylogeny was then constructed with RAxML [71] under the GTR+G evolutionary model.

Real-Time Quantitative RT-PCR

All samples were collected from 35 lab-reared colonies from clonal lineages MLL1, MLL4 and MLL6 [S2]. Colonies were inspected to assess their position in the colony cycle according to presence of eggs, larvae, prepupae or pupae, and eight one-month-old workers were collected at the appropriate stage of the colony cycle (worker age was assessed by the level of melanization [S31]). Workers collected from colonies in the brood care phase were only collected from the foraging arena. For all other cycle stages, workers were collected from within the nest. All workers from each colony were pooled for RNA extraction and further analysis. For head/abdominal expression, heads and abdomens were separately pooled from eight dissected workers from each colony.

Workers were placed in dry-ice-cooled ethanol upon collection, and maintained at -80 °C until processed. RNA was extracted using Trizol (Invitrogen) followed by RNeasy (Qiagen) purification, according to the manufacturers' recommended protocols. RNA was treated using DNAse I (Sigma), and quantified using a Nanodrop 2000. cDNA synthesis using polyT primers (Transcriptor First Strand Kit, Roche) was performed in duplicate for each sample, using 302 ng RNA.

Primer3 [S56] was used for primer design: Melt temperatures of each primer were set between 58-60°C, with a maximum difference between primer pairs of 1°C. PCR product length was set between 50 and 150 bp, with primer length between 18-22 bp. The last 5 bp of the 3' end of each primer were chosen to contain a maximum of two G/C bases. All primer pairs, where possible, span an intron. Amplicons in the first 1 kb of the cDNA sequence were preferred over more distal loci. Both primers and PCR products were filtered for closed secondary structures that would inhibit amplification efficiency.

All primers were used to amplify both cDNA and genomic template to confirm the presence of a single amplicon, and for intron-spanning pairs to show amplicon size difference between cDNA- and genomic DNA-based products. cDNA template was serially diluted with a total dilution factor of 512 across four concentrations, which were amplified in triplicate to assess amplification efficiency. Primers with efficiency between 90-110% and an $R^2 \geq 0.980$ were chosen for actual experiments:

Table of RT-qPCR primers used in experiments.

Gene	**Accession #**	**Fwd**	**Rev**
Actin	Cbir_12877	ATCCACGAGACCACGTACAA	TGATCTCCTTCTGCATCCTG
Tubulin α1b	Cbir_06820	TCGATTTGGAACCCACTGTA	CATAGTTGTTCGCAGCGTCT
CG13220	Cbir_11465	CATGAACCCAAGTCTTGTCG	TGTTCGCGCGTATAAAGGTA
Ef1α	Cbir_02512	GTTGGCTTCAACGTCAAGAA	CGGGATGATTGAGAACAATG
For	Cbir_15056	TATACGAGAGGGCGATGTTG	GGTGCCAAAGTGCTGAGATA
Gapdh1	Cbir_02225	GATCCACGACAACTTCGAGA	CGGCGGGAATAATGTTTT
Gst1	Cbir_01706	TGGCAAAAATGATTCCCTCT	TCGCCTGATCTGGAGTAGC
Hmbs	Cbir_05182	TGTATGGTCCCTTGACGGTA	AGGCGAGGTTCTTGATATGG
Hmgcr	Cbir_05631	TTCTGCAGCGTGGTTTTTAC	TGATTTTGCGAGCTTAGTGC
Mvl	Cbir_06948	ATTTGCTCACGGTCTCTTCC	CAGATCTGCGTTGAACGTCT
Rpl13α	Cbir_03893	GCAAACAAAAGCGTGTCAAG	TCGGCCAGAGTAAAACCTCT
Rpl32	Cbir_02279	CCGGTCTATCGACCAAAGAT	CCCTTAAAACGCCTACGAAC
Rps18	Cbir_05833	TGACTGCGATCAAAGGTGTT	TGGCCATAATGGTGACAATC
Rps3	Cbir_02386	AGCTATACGCGGAGAAGGTC	GCCCGATTCCATGATAAAAC
Rps6	Cbir_11616	TTTCCCATGAAGCAGGGTAT	GGATTTCTCCCTTGGACAAC
Sdhaα	Cbir_06668	ATGGTCTCCAGAGCCAACTT	GAATCTCTCGCCTTCGCTAT
Sdhaβ	Cbir_08924	ACATTTTCTTGCGAGGGAAG	CCGAAAGGACGCTGATAAAT
Syntaxin1	Cbir_11029	AGCGGTGTAAGGGGAGAATA	CTGTTTTGCTTGTTGCGTTT
Syntaxin5	Cbir_12585	CGCACTCCTCTTCGATTGTA	TGGTGGAGACAGATCCTTGA
Tbp	Cbir_03719	CTGCGAGAAAGTACGCAAGA	GAATTGTCCGTGAGAGAGCA
Vgq	Cbir_02775	ATCCGACTGCGAGTCTTCTT	GCCGAAGTAATCGTTGTTCA
Vgw	Cbir_06786	AAATGGTCGCATATGTCCAA	ACGTTTTATGGCTGGCTACC

To identify reference genes, *C. biroi* reciprocal best BLAST hits of *Rps3*, *Rps6*, *Rps18*, *Hmbs*, *Syntaxin1*, *Syntaxin5*, *Tbp*, *Sdhaα*, *Sdhaβ*, *Rpl13α*, *Gapdh1*, *Tubulin lpha 1B*, *Actin*, *Ef1α*, *Rpl32*, *CG13220*, and *Gst1* were assessed for variability between reproductive and brood care phases according to Vandesompele *et al*. [S57]. The three genes with the best expression stability values (M) were *Rps3*, *Rps6* and *Rpl13α*. These three genes were therefore chosen as reference genes in all analyses.

qPCR was performed using SYBR green (Maxima Master Mix; Thermo Fisher) on a Roche LightCycler 4.80, using the following amplification protocol: 1 cycle (8 min @ 95°C), 40 cycles (30s @ 95°C, 30s @ 60°C, 30s @ 72°C), followed by heat dissociation. All samples were amplified in triplicate, along with no-reverse-transcriptase controls and no-template controls. Gene expression data were normalized using the three reference genes according to Vandesompele *et al*. [S57].

RAD-Seq Sample Collection and Sequencing

Eight workers were collected from each of 15 live laboratory colonies in 2012: Five each from clonal lineages MLL1 (colonies were originally collected on Okinawa, Japan, in 2008), MLL4 (St. Croix, USA, 2010) and MLL6 (Okinawa, Japan, 2008). Additionally, eight workers each from four colonies of MLL13, which is a new clonal lineage that has not been studied previously, were collected directly into 96% ethanol in June 2011 in Lianhuashan Park, Shenzhen, China. Approximately 0.5 μg of DNA was extracted from each worker using a phenol-chloroform protocol. DNA was treated with restriction enzyme EcoRI for library construction.

DNA was digested with EcoRI, adapter ligated and sequenced on an Illumina HiSeq 2000 to a miimum genome-wide depth of 1.3x. Reads were aligned to the *C. biroi* genome using BWA [S58] and SNPs identified and filtered using SAMtools [S59] (default parameters).

RAD-Seq Analysis

Because *C. biroi* reproduces asexually via automixis with central fusion, variation between individuals within a given clonal lineage will be predominantly the result of losses of heterozygosity (LOH). By comparing the genotypes of individuals from the same clonal lineage, it is therefore possible to reconstruct the ancestral genotype of those samples: A locus with two alleles in the clonal lineage, even if exclusively homozygous in any given individual, is most parsimoniously explained as originating from an ancestrally heterozygous locus. For each clonal lineage, we therefore measured the frequency of all alleles present in our samples, and considered the ancestral genotype to consist of the two most frequent alleles (with a minimum read depth of 20 for each allele) [S60]. This method has the advantage of minimizing the signal of genotyping errors, and provides a cumulative read depth high enough to sample both alleles at ancestrally heterozygous loci.

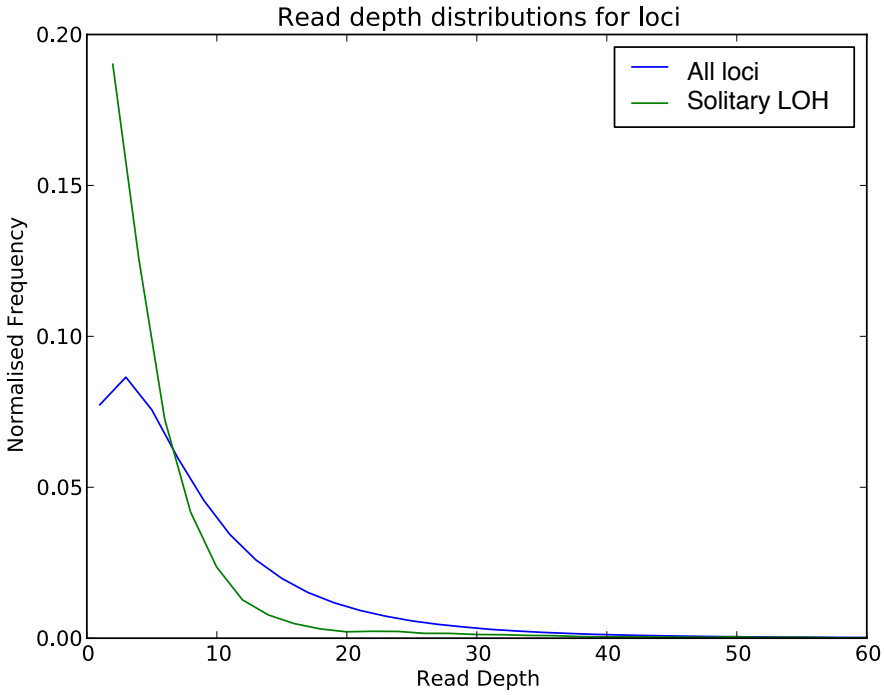
During sequencing, the sampling of alleles at a given locus is random. If the read depth at a locus is low, there is a high probability that only one allele will be sampled. This will spuriously increase the number of homozygous loci observed for a given individual. This is particularly problematic for RAD-Seq analyses of clonal organisms. Because most variation between individuals is expected to arise from LOH, allelic sampling bias can contribute to a significant portion of the apparent differences between individuals.

In our *C. biroi* samples, the average read depth for all loci observed using RAD-Seq was 7x. Using a Poisson sampling distribution of read depth [S61], and a binomial distribution of allele sampling for each allele of a specified read depth, it is possible to calculate the expected number of type II errors in each clonal lineage. Increasing the read depth required at a given locus will reduce the type II error rate, but at a cost of simultaneously reducing the number of loci available for analysis.

Because *C. biroi* reproduces through automixis with central fusion, most LOH will occur as a consequence of meiotic crossover events. These will result in large regions

of the genome becoming homozygous. We therefore expect most loci with LOH to be adjacent to other ancestrally heterozygous loci that have experienced LOH. Furthermore, in asexual populations, homozygosity will spread through the population over time as it is inherited from individual to individual. We therefore expect most LOH events to be shared by more than one individual in each clonal lineage sampled, especially because our sampling regime included individuals from the same colony [S62]. 'Solitary' LOH events, i.e. those observed in only a single individual or covering a single locus, are therefore more likely to be a result of type II errors.

Indeed, by comparing the read depth of all loci to the read depth of solitary LOH loci, we found a bias among solitary LOH loci for low read depth:



To reduce the type II error rate in our analysis, we therefore employed a read depth threshold conditional on the locus genotype. Individual genotypes that differed from the ancestral genotype required a minimum read depth of 15 in order to be scored.

An additional source of genotyping bias is introduced during library construction. This can result in only a single allele of a heterozygous locus being represented in the library, which will not be corrected by increasing the read depth. To assess the level of this bias in our samples, we compared 17 individuals for which two libraries were independently constructed and sequenced. Loci with a read depth ≥ 15 were identified that were homozygous in one library but heterozygous in the corresponding library. Using these known false homozygous loci and the number of loci homozygous in both libraries (assumed to be true homozygous loci), we found an average false homozygous error rate of 0.5% of the loci scored as homozygous. This corresponds to approximately 3.5 times the frequency of solitary LOH loci identified in any given

individual. Therefore, based on the expected mechanism for LOH under automixis with central fusion, the expectation for recurrent genotypes in an asexual population, and the observed library error rates in our samples, we excluded all solitary LOH loci from our analysis. The resulting filtered data were used in all subsequent analyses – calculating relatedness, phylogeny, and LOH number and size.

To validate our RAD-Seq data, we compared the LOH observed at previously published microsatellite loci [S10] with the LOH in our RAD-Seq SNPs 10 kb upstream and 10 kb downstream of those microsatellite loci. In all cases where microsatellites had lost heterozygosity in a colony [S10], the RAD-Seq data also showed LOH of ancestrally heterozygous SNPs.

Because this is the first RAD-Seq dataset used to calculate LOH in an asexual species, comparisons with the rates determined through microsatellite analysis (such as for the Cape honey bee) may not correspond exactly. Variability in the number and genomic position of microsatellites used in other studies will affect the LOH rate observed, and estimates based on small numbers of markers might be especially imprecise. This is particularly the case for species that reproduce via automixis, as distance from the centromeres will determine the likelihood of recombination and LOH. However, the Cape honey bee LOH rate with which we compare our estimates for *C. biroi* was determined based on a large number (161) of microsatellite loci. These loci were evenly distributed across the entire genome, in order to avoid the bias associated with distance from the centromeres. We therefore believe that our comparison of rate estimates for the Cape honey bee and *C. biroi* are valid.

It is also important to note that the exact ages of the *C. biroi* clonal lineages are unknown, and it is therefore only possible to determine the maximum LOH rate for each clonal lineage, based on the date of earliest collection. Given that *C. biroi* has been established invasively for over 100 years, the actual LOH rate may be significantly slower than our current estimates.

Table showing LOH rates obtained with and without exclusion of solitary LOH loci. Even without the exclusion of solitary LOH loci, the rate of LOH is as low as 0.0025% per generation: 51.7 fold lower than in *A. mellifera capensis*.

Clonal lineage	**Number of generations since field collection**	**LOH rate with all loci (LOH per generation)**	**LOH rate without solitary LOH loci (LOH per generation)**
MLL1	225	2.5×10^{-5}	1.3×10^{-5}
MLL4	86	6.3×10^{-5}	3.5×10^{-5}
MLL6	43	6.2×10^{-5}	2.3×10^{-5}
MLL13	0	2.0×10^{-3}	1.0×10^{-3}

DNA Methylation and Histone Modification

InterPro domains from the functional annotation were used to identify all genes associated with DNA and histone modifications. Genes from *Apis mellifera* involved in DNA methylation were additionally used to search the *C. biroi* genome using tBLASTn to check for potential paralogs. Like the other sequenced ant species, *C. biroi* has the full complement of DNA methylation enzymes. Because methylated cytosine mutates more readily to uracil than unmethylated cytosine, genes that are methylated in the germline have a lower observed CpG ratio than expected based on nucleotide frequency. The $CpG_{(observed/expected)}$ distribution was calculated for the coding sequences of all sequenced ants, *D. melanogaster, A. pisum,* and *A. mellifera* (Figure S1). The hypergeometric test to assess whether genes with a $CpG_{(obs/exp)}$ ratio less than 1 were enriched for any GO categories did not identify any significantly enriched GO terms in *C. biroi*.

Table of genes with domains for DNA and RNA methylation, chromatin and histone modification in *C. biroi*.

Accession #	Gene Name
IPR000286; Histone deacetylase superfamily	
Cbir_03995	similar to Hdac3
Cbir_05949	similar to Rpd3, Cbir paralog 1
Cbir_07104	similar to Rpd3, Cbir paralog 2
Cbir_07204	similar to HDAC6
Cbir_14891	similar to HDAC4
IPR000953; Chromo domain/shadow	
Cbir_05703	similar to Chd3
Cbir_04638	similar to kis
Cbir_07228	similar to Chd1
Cbir_10412	
Cbir_03016	similar to Chro
Cbir_03702	similar to HP1b
Cbir_04240	similar to mof
Cbir_05269	similar to Su(var)3-9
Cbir_06355	
Cbir_06466	
Cbir_06735	similar to MRG15
Cbir_06799	
Cbir_06808	
Cbir_08160	
Cbir_11040	
Cbir_11137	similar to msl-3
Cbir_11688	
Cbir_14678	

Accession #	Gene Name
IPR001214; SET domain	
Cbir_02709	similar to egg
Cbir_08659	
Cbir_07175	similar to Mes-4
Cbir_01725	similar to CG5591
IPR001214; SET domain	
Cbir_06469	similar to CG40351
Cbir_04238	
Cbir_03119	similar to G9a
Cbir_03664	similar to CG9642
Cbir_04180	similar to CG8378
Cbir_04211	
Cbir_04956	
Cbir_05033	
Cbir_05259	
Cbir_05470	similar to Blimp-1
Cbir_05521	similar to E(z)
Cbir_05667	
Cbir_05921	similar to CG3353
Cbir_06675	similar to pr-set7
IPR001214; SET domain	
Cbir_08626	
Cbir_09928	similar to msta
Cbir_10630	similar to CG14590
Cbir_12393	similar to CG7759
Cbir_14116	similar to CG4565
Cbir_14479	
Cbir_14638	similar to ham
Cbir_15299	similar to CG32732

Table of genes with domains for DNA and RNA methylation, chromatin and histone modification in *C. biroi*. (continued)

Accession #	Gene Name
IPR001487; Bromodomain	
Cbir_03303	similar to bon
Cbir_02251	similar to Pcaf, Cbir paralog 1
Cbir_06660	similar to Pcaf, Cbir paralog 2
Cbir_05260	similar to nej
Cbir_03354	similar to Br140
Cbir_06494	similar to CG1815
Cbir_06763	
Cbir_05561	similar to brm
Cbir_04026	similar to polybromo
Cbir_02086	similar to Acf1
Cbir_02651	similar to Taf1
Cbir_05335	similar to E(bx)
Cbir_05548	similar to dikar
Cbir_05627	similar to CG7154
Cbir_06583	
Cbir_06665	similar to Brd8
Cbir_10077	
Cbir_12476	
Cbir_14655	similar to tou
Cbir_15389	similar to BRWD3
IPR001525; C-5 cytosine methyltransferase	
Cbir_06841	similar to DNMT1
Cbir_05899	similar to TRDMT1
Cbir_00054	similar to DNMT3
IPR001739; Methyl-CpG DNA binding	
Cbir_03315	
Cbir_03602	similar to MBD-R2 (MECP2)
Cbir_03943	similar to MBD-like
Cbir_16323	
IPR001965; Zinc finger, PHD-type	
Cbir_05440	similar to lid
Cbir_03311	similar to CG3815
Cbir_05169	
Cbir_05108	similar to CG2662
Cbir_01936	
Cbir_06628	similar to CG9576
IPR001965; Zinc finger, PHD-type	
Cbir_06879	similar to ash2
Cbir_02562	
Cbir_03225	similar to CG7379
Cbir_03560	
Cbir_04098	similar to IntS12
Cbir_04343	
Cbir_04481	

Accession #	Gene Name
IPR001965; Zinc finger, PHD-type	
Cbir_04523	similar to enok
Cbir_05877	similar to CG15439
Cbir_05928	similar to Kdm2
Cbir_06313	similar to Kdm4B
Cbir_07885	similar to Ing3
Cbir_09841	similar to CG17446
Cbir_09951	
Cbir_10608	
Cbir_10784	
Cbir_13640	
Cbir_13765	
Cbir_13945	
Cbir_13959	similar to Pcl
Cbir_15251	similar to d4
IPR003000; NAD-dependent deacetylase, sirtuin family	
Cbir_07069	similar to Sirt4
Cbir_07211	similar to Sirt2
Cbir_08551	similar to Sirt7
Cbir_09114	
Cbir_13613	similar to Sirt6
Cbir_14028	
IPR003347; JmjC domain	
Cbir_15136	similar to Utx
Cbir_02800	
Cbir_02228	
Cbir_02795	similar to CG2982
Cbir_04634	similar to JHDM2
Cbir_05233	
Cbir_09300	similar to CG10133
Cbir_10959	similar to PSR
Cbir_13468	similar to CG7200
Cbir_15606	similar to CG13902
Accession #	**Gene Name**
IPR004092; Mbt repeat	
Cbir_02163	similar to Scm
Cbir_05758	similar to l(3)mbt
Cbir_06999	similar to Sfmbt
Cbir_05959	
IPR007526; SWIRM domain	
Cbir_02461	similar to mor
Cbir_10941	similar to Rpb4
Cbir_01669	similar to Su(var)3-3
Histone Phosphorylation	
Cbir_03604	similar to Haspin
Cbir_12827	similar to CkIIalpha

miRNAs

Identified pre-miRNAs from model organisms (mirbase.org, release 19) were used as query sequences for a BLASTN search (default parameters, E-value < 1.8) against the *C. biroi* genome. The mature miRNAs were then aligned to the *C. biroi* target sequences, and only miRNAs with less than four nucleotide mismatches were retained [S63]. This method revealed 63 miRNAs in *C. biroi*:

Table of miRNAs identified in the *C. biroi* genome.

miRNA	Scaffold	Position on scaffold
bantam	scaffold544	1392379-1392451
let-7	scaffold101	289179-289087
mir-1	scaffold273	567483-567456
mir-10	scaffold334	339661-339586
mir-100	scaffold101	290003-289909
mir-1000	scaffold236	633849-633935
mir-11	scaffold910	217624-217520
mir-12	scaffold4	1900559-1900490
mir-124	scaffold3	1161755-1161784
mir-125	scaffold101	288411-288305
mir-133	scaffold273	443161-443134
mir-137	scaffold137	118232-118322
mir-13a	scaffold903	288025-288122
mir-13b	scaffold903	288392-288470
mir-14	scaffold59	430579-430495
mir-184	scaffold581	172419-172492
mir-193	scaffold59	843853-843769
mir-210	scaffold436	737339-737421
mir-219	scaffold1503	718-814
mir-252	scaffold74	433508-433420
mir-252a	scaffold74	433508-433414
mir-263a	scaffold800	840779-840867
mir-263b	scaffold105	201491-201578
mir-275	scaffold411	808329-808417
mir-276	scaffold538	63947-63864
mir-277	scaffold390	215388-215471
mir-278	scaffold429	1995297-1995370
mir-2796	scaffold260	3340936-3341024

miRNA	Scaffold	Position on scaffold
mir-281	scaffold136	1392947-1393046
mir-282	scaffold94	61822-61907
mir-283	scaffold4	1902060-1901962
mir-2b	scaffold903	288608-288670
mir-305	scaffold411	808540-808627
mir-307	scaffold429	1807346-1807253
mir-315	scaffold606	387156-387079
mir-316	scaffold125	521250-521336
mir-317	scaffold390	197428-197517
mir-31a	scaffold260	1531168-1531092
mir-33	scaffold427	108165-108251
mir-34	scaffold390	218552-218642
mir-3477	scaffold4	1901010-1900919
mir-375	scaffold150	1464208-1464268
mir-3777	scaffold429	187247-187147
mir-3783	scaffold346	1833067-1833150
mir-3786	scaffold16	526403-526319
mir-6012	scaffold103	213776-213693
mir-6038	scaffold436	731939-732011
mir-6067	scaffold16	492976-492899
mir-7	scaffold20	686585-686655
mir-71	scaffold903	287417-287499
mir-750	scaffold133	189759-189681
mir-8	scaffold411	368981-368910
mir-927	scaffold665	557629-557701
mir-927a	scaffold665	557628-557722
mir-927b	scaffold260	1314583-1314491
mir-929	scaffold74	489146-489052
mir-92a	scaffold314	721909-721826
mir-932	scaffold113	2095200-2095290
mir-980	scaffold101	331732-331803
mir-993	scaffold334	299336-299423
mir-9a	scaffold399	114756-114692
mir-iab-4	scaffold334	857409-857493
mir-iab-8	scaffold334	857483-857419

Repeats

RepeatModeler [S64] was used to construct a *C. biroi*-specific repeat library, which was used by RepeatMasker [S65] to identify repeats in the genome.

Table showing repeat elements identified in the *C. biroi* genome.

Repeat element	**Number of elements**	**Total length occupied (bp)**	**Percentage of sequence**
SINES	249	41,144	0.02%
LINES	3,994	1,471,116	0.69%
LTR elements	1,329	987,711	0.46%
DNA elements	23,259	7,948,856	3.71%
Unclassified	59,782	13,812,226	6.44%
Small RNAs	112	16,812	0.01%
Satellites	281	57,298	0.03%
Simple repeats	90,533	4,358,631	2.03%
Low complexity	15,364	810,137	0.38%

***Transformer* Genes**

Cerapachys biroi was searched using TBLASTN with the other ant *CSD* and *Feminizer* homologs as search queries. This revealed only a single homolog for *Feminizer* and none for *CSD*. *Feminizer* and *CSD* are adjacent or separated by only a few genes in all the other sequenced ant species and *A. mellifera*, and in ants, *CSD* and *Feminizer* show inter-locus recombination and conserved synteny [S66]. It is therefore expected that, if present, *CSD* would be located near *Feminizer* in *C. biroi*. The *Feminizer* homolog identified in *C. biroi* is located in the center of a large scaffold with no large assembly gaps in which *CSD* may be located. Additionally, the genomic region in *C. biroi* showed synteny with the scaffolds in the other sequenced ants, both up- and downstream of *Feminizer*.

Supplemental References

S1. Smith, C., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C., Elhaik, E., Elsik, C., et al. (2011). Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proc. Natl. Acad. Sci. USA *108*, 5673-5678.

S2. Smith, C., Smith, C., Robertson, H., Helmkampf, M., Zimin, A., Yandell, M., Holt, C., Hu, H., Abouheif, E., Benton, R., et al. (2011). Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc. Natl. Acad. Sci. USA *108*, 5667-5672.

S3. Zhou, X., Slone, J., Rokas, A., Berger, S., Liebig, J., Ray, A., Reinberg, D., and Zwiebel, L. (2012). Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. PLoS Genet. *8*, e1002930.

S4. Vieira, F.G., Forêt, S., He, X., Rozas, J., Field, L.M., Zhou, J.-J., and Robinson-Rechavi, M. (2012). Unique features of odorant-binding proteins of the parasitoid wasp *Nasonia vitripennis* revealed by genome annotation and comparative analyses. PLoS ONE *7*, e43034.

S5. Kulmuni, J., Wurm, Y., and Pamilo, P. (2013). Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. Heredity *110*, 538-547.

S6. Robertson, H., and Wanner, K. (2006). The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res. *16*, 1395-1403.

S7. Forêt, S., and Maleszka, R. (2006). Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). Genome Res. *16*, 1404-1413.

S8. Forêt, S., Wanner, K., and Maleszka, R. (2007). Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. Insect Biochem. Mol. Biol. *37*, 19-28.

S9. Khila, A., and Abouheif, E. (2009). *In situ* hybridization on ant ovaries and embryos. Cold Spring Harb. Protoc., http://dx.doi.org/10.1101/pdb.prot5244.

S10. Kronauer, D.J.C., Pierce, N., and Keller, L. (2012). Asexual reproduction in introduced and native populations of the ant *Cerapachys biroi*. Mol. Ecol. *21*, 5221-5235.

S11. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265-272.

S12. Elsik, C., Mackey, A., Reese, J., Milshina, N., Roos, D., and Weinstock, G. (2007). Creating a honey bee consensus gene set. Genome Biol. *8*, R13.

S13. Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061-1067.

S14. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.

S15. Cantarel, B., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188-196.

S16. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

S17. Slater, G., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

S18. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988-995.

S19. Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. *32*, W309-W312.

S20. Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511-515.

S21. Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. *37*, 289-297.

S22. Bairoch, A. (2004). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154-D159.

S23. Zdobnov, E., and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

S24. Schug, J., Diskin, S., Mazzarelli, J., Brunk, B., and Stoeckert, C. (2002). Predicting gene ontology functions from ProDom and CDD protein domains. Genome Res. *12*, 648-655.

S25. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35*, W182-W185.

S26. Li, L., Stoeckert, C., and Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178-2189.

S27. De Bie, T., Cristianini, N., Demuth, J., and Hahn, M. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269-1271.

S28. Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792-1797.

S29. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696-704.

S30. Lee, E., Harris, N., Gibson, M., Chetty, R., and Lewis, S. (2009). Apollo: a community resource for genome annotation editing. Bioinformatics *25*, 1836-1837.

S31. Teseo, S., Kronauer, D.J.C., Jaisson, P., and Châline, N. (2013). Enforcement of reproductive synchrony via policing in a clonal ant. Curr. Biol. *23*, 328-332.

S32. Ravary, F., Jahyny, B., and Jaisson, P. (2006). Brood stimulation controls the phasic reproductive cycle of the parthenogenetic ant *Cerapachys biroi*. Insectes Soc. *53*, 20-26.

S33. Maleszka, J., Forêt, S., Saint, R., and Maleszka, R. (2007). RNAi-induced phenotypes suggest a novel role for a chemosensory protein CSP5 in the development of embryonic integument in the honeybee (*Apis mellifera*). Dev. Genes Evol. *217*, 189-196.

S34. Nomura, A., Kawasaki, K., Kubo, T., and Natori, S. (1992). Purification and localization of p10, a novel protein that increases in nymphal regenerating legs of *Periplaneta americana* (American cockroach). Int. J. Dev. Biol. *36*, 391-398.

S35. Leal, W., Ishida, Y., and El-Shemy, H.A. (2008). GP-9s are ubiquitous proteins unlikely involved in olfactory mediation of social organization in the red imported fire ant, *Solenopsis invicta*. PLoS ONE *3*, e3762.

S36. Park, S.K., Shanbhag, S.R., Wang, Q., Hasan, G., Steinbrecht, R.A., and Pikielny, C.W. (2000). Expression patterns of two putative odorant-binding proteins in the olfactory organs of *Drosophila melanogaster* have different implications for their functions. Cell Tissue Res. *300*, 181-192.

S37. Pelosi, P., Zhou, J.J., Ban, L., and Calvello, M. (2006). Soluble proteins in insect chemical communication. Cell. Mol. Life Sci. *63*, 1658-1676.

S38. Leal, W. (2013). Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. Annu. Rev. Entomol. *58*, 373-391.

S39. Real, M., Ferré, J., and Chapa, F. (1991). UDP-glucosyltransferase activity toward exogenous substrates in *Drosophila melanogaster*. Anal. Biochem. *194*, 349-352.

S40. Bull, D., and Whitten, C. (1972). Factors influencing organophosphorus insecticide resistance in tobacco budworms. J. Agric. Food Chem. *20*, 561-564.

S41. Lazard, D., Zupko, K., Poria, Y., Nef, P., Lazarovits, J., Horn, S., Khen, M., and Lancet, D. (1991). Odorant signal termination by olfactory UDP glucuronosyl transferase. Nature *349*, 790-793.

S42. Wang, Q., Hasan, G., and Pikielny, C. (1999). Preferential expression of biotransformation enzymes in the olfactory organs of *Drosophila melanogaster*, the antennae. J. Biol. Chem. *274*, 10309-10315.

S43. Mackenzie, P., Owens, I., Burchell, B., Bock, K., Bairoch, A., Bélanger, A., Fournel-Gigleux, S., Green, M., Hum, D., Iyanagi, T., et al. (1997). The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. Pharmacogenetics *7*, 255-269.

S44. Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673-4680.

S45. Ahn, S.-J., Vogel, H., and Heckel, D. (2012). Comparative analysis of the UDP-glycosyltransferase multigene family in insects. Insect Biochem. Mol. Biol. *42*, 133-147.

S46. Luque, T., Okano, K., and O'Reilly, D. (2002). Characterization of a novel silkworm (*Bombyx mori*) phenol UDP-glycosyltransferase. Eur. J. Biochem. *269*, 819-825.

S47. Lazard, D., Tal, N., Rubinstein, M., Khen, M., Lancet, D., and Zupko, K. (1990). Identification and biochemical analysis of novel olfactory-specific cytochrome P-450IIA and UDP-glucuronosyl transferase. Biochemistry *29*, 7433-7440.

S48. Huang, F.-F., Chai, C.-L., Zhang, Z., Liu, Z.-H., Dai, F.-Y., Lu, C., and Xiang, Z.-H. (2008). The UDP-glucosyltransferase multigene family in *Bombyx mori*. BMC Genomics *9*, 563.

S49. Meech, R., and Mackenzie, P. (1997). Structure and function of uridine diphosphate glucuronosyltransferases. Clin. Exp. Pharmacol. Physiol. *24*, 907-915.

S50. Burchell, B., and Coughtrie, M. (1989). UDP-glucuronosyltransferases. Pharmacol. Ther. *43*, 261-289.

S51. Hopkins, T.L., and Kramer, K.J. (1992). Insect cuticle sclerotization. Annu. Rev. Entomol. *37*, 273-302.

S52. Daimon, T., Hirayama, C., Kanai, M., Ruike, Y., Meng, Y., Kosegawa, E., Nakamura, M., Tsujimoto, G., Katsuma, S., and Shimada, T. (2010). The silkworm *Green b* locus encodes a quercetin 5-O-glucosyltransferase that produces green cocoons with UV-shielding properties. Proc. Natl. Acad. Sci. USA *107*, 11471-11476.

S53. Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680-682.

S54. Nelson, D. (2006). Cytochrome P450 nomenclature, 2004. Methods Mol. Biol. *320*, 1-10.

S55. Feyereisen, R. (2006). Evolution of insect P450. Biochem. Soc. Trans. *34*, 1252-1255.

S56. Rozen, S., and Skaletsky, H.J. (1998). Primer3. http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

S57. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. *3*, 1-12.

S58 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics *25*, 1754-1760.

S59 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (SAM) format and SAMtools. Bioinformatics *25*, 2078-2079

S60. Wang, Y., Lu, J., Yu, J., Gibbs, R.A., and Yu, F. (2013). An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. *23*, 833-842.

S61. Wheeler, D., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872-876.

S62. King, C.E., and Schonfeld, J. (2001). The approach to equilibrium of multilocus genotype diversity under clonal selection and cyclical parthenogenesis. Hydrobiologia *446*, 323-331.

S63. Lim, L., Lau, N., Weinstein, E., Abdelhakim, A., Yekta, S., Rhoades, M., Burge, C., and Bartel, D. (2003). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991-1008.

S64. Smit, A.F.A., and Hubley, R. (2008-2010). RepeatModeler Open-1.0. http://www.repeatmasker.org.

S65. Smit, A.F.A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0. http://www.repeatmasker.org.

S66. Privman, E., Wurm, Y., and Keller, L. (2013). Duplication and concerted evolution in a master sex determiner under balancing selection. Proc. R. Soc. Lond. B Biol. Sci. *280*, 20122968.